DOCUMENT RESUME

ED 069 747 TM 002 202

AUTHOR Cameron, Bernard J.; And Others
TITLE Operational Evaluation from the Standpoint of the Program Manager.
INSTITUTION BioTechnology, Inc., Arlington, Va.
SPONS AGENCY Bureau of Elementary and Secondary Education (DHEW/OE), Washington, D.C.
PUB DATE Oct 71
CONTRACT OEC-0-70-4951(284)
NOTE 48p.; This is the second of two documents prepared under the contract

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Costs; Evaluation Criteria; Evaluation Methods; *Evaluation Techniques; Measurement Instruments; *Operations Research; *Program Evaluation; *Research Methodology; Technical Reports; Test Construction
IDENTIFIERS Belmont Training Programs

ABSTRACT

The limits, function and procedures of operational evaluation are described. Operational evaluation can only begin once a project activity is underway. Its function is diagnostic but not prescriptive. Basic tasks include specifying objectives, defining criteria, establishing priorities, identifying cost factors, obtaining or developing measurement procedures and tools, and providing techniques to measure side effects. Types of analysis described are means, constraints, formulative, and summative. Effort, efficiency and effectiveness may be evaluated. The Belmont training programs are used to illustrate operational procedures. A section on methodology describes the development of instruments and design tactics. The final section deals with a consideration of problems related to the personnel who conduct operational studies. (DJ)

PPE
TM

OPERATIONAL EVALUATION FROM THE STANDPOINT OF THE PROGRAM MANAGER

Bernard J. Cameron
Jerry S. Kidd
Harold E. Price

October 1971

Prepared under Contract OEC–0–70–4951 (284) for

United States Office of Education
Bureau of Elementary and Secondary Education
Office of Program Planning & Evaluation

BioTechnology, Inc.
3027 ROSEMARY LANE · FALLS CHURCH, VIRGINIA

OPERATIONAL EVALUATION FROM THE STANDPOINT OF THE PROGRAM MANAGER

Bernard J. Cameron
Jerry S. Kidd
Harold E. Price

October 1971

Prepared under Contract OEC-0-70-4951(284) for

United States Office of Education
Bureau of Elementary and Secondary Education
Office of Program Planning & Evaluation

FOREWORD

This report, Operational Evaluation from the Standpoint of the Program Manager was prepared for the U.S. Office of Education by BioTechnology, Inc. under Contract OEC–0–70–4951(284). It is the second of two documents prepared under that contract. The first report, An Evaluation of the Belmont Training Program was produced in January, 1971.

TABLE OF CONTENTS

Operational Evaluation from the Standpoint of the Program Manager

Introduction

In the ironic phraseology of the old Chinese curse, we are living in an "interesting time." The field of public administration is characterized by problems that, more often than not, seem to escalcte rapidly to crisis proportions. The conditions within which programs are to be conducted change from day to day. The problems that programs and projects are intended to solve also emerge with such unprecedented attributes that established operational methods and procedures are made obsolete. Program administrators and project managers are continually faced with the demand to invent new ways of operating. They must assemble resources and specify procedures for doing some job that has never been done before.

Inevitably, in these circumstances, some of these improvised processes work better than others and some do not work at all. One final complication, however, is that it is often difficult to tell whether one new way of doing something is good, bad, or indifferent. Success in public program operations is often hard to define and even harder to measure. This situation has led to the self-conscious development of techniques for operational evaluation. That is, there has come into being a new vocational specialty based on the analysis and evaluation of projects and programs in the domain of public administration.* The logic is that if new operational methods must be used in situations where failure could have disasterous consequences, some device or process is needed to forestall failure, weed out deficiencies, and ensure reasonable economy, while the whole process is going on.

While a body of doctrine and a group of specialists now exists, operational evaluation is still not a very mature field in the sense that its structure is comprised of a hodge-podge of fragments; many of these fragments have merit on their own but they do not automatically fit together into a coherent array. The lines of fragmentation follow boundaries between antecedent disciplines (e.g., accounting,

*The emergence of a new field of specialization is symptomized by the establishment of a professional organization, the Association of Public Program Analysis, in 1966; which numbers as members mostly government officials plus a few academics with close ties to governmental operations.

the engineering sciences, and the social sciences) and the operational domains (e.g., education, public health, urban affairs, welfare, military technology, etc.). Thus, while there are tools and techniques which probably have broad if not universal applicability, one is best advised to take a rather cautious approach to each new evaluational task with the prospect in mind that what was appropriate last time might be inappropriate now and that some cutting and fitting to meet the specific conditions is probably required.

What is generalizable is the basic framework of the problem from the administrator's or manager's point of view. Or, at least this framework can be made general by careful pruning.

To get at this general framework, we first must be able to epitomize one of the major task areas in the work of the public program administrator and project manager. Very simply, the task area with which we are concerned is that of problem solving. As defined here, this involves the putting together of resources, capabilities, and procedures in an attempt to accomplish some objective. We are *not* concerned with policy making, the specification of objectives, the initial acquisition of resources, program staffing, human relations in management, or any other of a host of things that can make up the complete work responsibilities of a program or project official. We are concerned rather with the tactical level of *how* the program, project, or subproject gets done, and whether it is done well and efficiently. We also are only tangentially concerned with alternative approaches. That is, it is not a central function of operational evaluation to yield still newer and better ways of doing a particular job; however desirable it might be to have such "corrective feedback" and irrespective of the likelihood that the results of operational evaluation can contain crucial clues for further development. In fact, the function of operational evaluation is (or should be) restricted to only one categorical decision in the problem solving process of program and project management. It is (or should be) strictly a *post hoc* decision: after the fact of putting a project or subproject into operation.

Thus, operational evaluation really can only begin once a delimitable project activity is underway. The decision framework which the outcome of operation evaluation can effect is simply:

Given an ongoing activity, (1) should it be continued as is, (2) should it be revised or adjusted, or (3) should it be discontinued? It should be obvious to anyone with experience in program

administration and project management that this decision framework is somewhat constrained at the outset—that is, before any operational evaluation findings are available. For example, management lore would have it that there is no such thing as an activity that cannot be improved. Moreover, unless the objective that the activity was meant to achieve can be eliminated, complete discontinuation of the activity will have low feasibility. Only a demonstration that the activity contributes nothing or is a positive impairment to the achievement of project or program objectives would permit cessation without a substitute activity to replace it. Finally, there is always the factor of inertia and a commitment to a way of doing something as it is being done a particular way by the people directly involved.

All these constraints and other factors as well contribute to one preeminent response to operational evaluation which is, in a phrase, marginal adjustment. That is, a wide range of outcomes of operational evaluation are likely to have only a narrow range of effects on the operational activity, no matter how rigorously or with what level of sophistication the evaluation work is done. This condition makes for a certain tendency toward cynicism among those who carry out operational evaluations and some misunderstanding of the whole process when it is viewed from a political or public policy standpoint.

To affirm the value of operational evaluation, we must go back to examining the "corrective feedback" notion and enlarge our consideration of it. Ideally, again, the function of operational evaluation is essentially diagnostic but not prescriptive. However, no prescription can be intelligently made without that prior diagnostic step and, more often than not, prescription is made much easier if the diagnosis is accurate.

The main point is that there is no other way to cope with the need. The syllogism goes like this:

- New and changing conditions impose objectives and there is a demand for action to meet those objectives.

- The objectives cannot often be met by the use of traditional or "tried and true" procedures.

- New procedures must be invented and some of these inventions do not work well.

- It saves money and frustration if mistakes are detected early and accurately.

While the knowledge that something is wrong does not tell you what to do about it, it does tell you what to do next, namely, find a way to do it better.

Operational evaluation, within this framework is recognized as being only a part of the job of operational problem solving; and operational problem solving is recognized as being only a part of the job of program administration and project management. It seems likely that past failures to make these fine distinctions have led to misunderstanding of the function and utility of operational evaluation, not only by outsiders, and not only by the administrative beneficiaries, but also by the practitioners themselves.

The Nature of Evaluation

Background

Evaluation seems to be as natural to human beings as breathing. In every day affairs, our first reaction to any novel experience tends to be evaluative: we either like it or do not like it. Individuals assigned the task of "discussing" any subject almost invariably begin their commentary with some form of evaluative statement.

While the evaluative response is extremely pervasive and presumably well practiced, it is not, under ordinary circumstances, particularly systematic. In every day affairs, we are dealing primarily with the phenomenon of individual tastes and preferences which are notoriously illogical, inconsistent, and variant from person to person. People consider things good or bad for obscure reasons and associate themselves with processes that often seem to the detached observer to be irrelevant or even destructive (e.g., cigarette smoking).

Throughout most of human history, the conduct of community business was virtually as casual as the conduct of one's private life. That is, judgments were made about the "goodness" of alternative policies and procedures on the basis of taste—the personal preference of the decision maker constrained by tradition. Only recently has there been any concerted attempt to approach the determination of public policies and program procedures in an orderly and rational manner.

It is noteworthy that the real beginnings of what might be called scientific evaluation of public policies and programs took place in the area of public health around the beginning of the century. Several factors probably contributed to this development. These included the magnitude and the visibility of such programs to important political constituencies, the clarity and measurability of the crucial criteria (i.e., mortality and morbidity rates) and the ingrained scientific orientation of the participants.

During the past fifty years, considerable maturation has taken place. The major milestone was World War II during which a convergence of scientific method and public policy was forged with

respect to military operations. The success of this amalgam was sufficiently apparent to both politicians and administrators that the concept of scientific participation has since been extended to all areas of public policy and program development. Moreover, the expectation, not to say the requirement, for operational evaluation is being extended across the complete hcirarchy of operations from component activities to complete programs. In some cases, each aspect of a major program may be evaluated independently while a summary evaluation is being undertaken for the program as a whole.

Evaluation Defined

The term evaluation is frequently used as if there were unanimous agreement about its meaning, and as if some common and unitary conceptual framework and methodology existed for evaluating any identifiable program. Possibly because of the pervasiveness of the evaluative response, however, the term evaluation is highly ambiguous. The technical literature abounds with a mixture of conceptual and operational definitions that is confusing even to those who are reasonably well informed in the area. (See, for example, Suchman, 1967, p. 28.) Any clear examination of evaluative activity reveals, moreover, not only highly divergent frames of reference but a wide range of methodological practices depending on the evaluative context, the complexity of the program, and its stage of development.

Perhaps the most precise, and at the same time, comprehensive definition of evaluation was recently formulated by Stufflebeam (1968). The definition is reproduced below.

> Generally, evaluation means the provision of information through formal means, such as criteria, measurement, and statistics, to serve as rational bases for making judgments in decision situations. To clarify this definition, it will be useful to define several key terms. A decision is a choice among alternatives. A decision situation is a set of alternatives. Judgment is the assignment of values to alternatives. A criterion is a rule by which values are assigned to alternatives, and optimally such a rule includes the specification of variables for

> measurement and standards for use in judging that which is measured. Statistics is the science of analyzing and interpreting sets of measurements. And, measurement is the assignment of numerals to entities according to rules, and such rules usually include the specification of sample elements, measuring devices and conditions for administering and scoring the measuring devices.

Stufflebeam goes on to indicate that the basic purpose of evaluation is to provide information upon which decisions can be based, to elaborate upon the methodological functions within the evaluation process, and to develop a taxonomy of educational decision making based on its function. For present purposes, however, we can settle for a relatively simple definition—evaluation is the process of determining the results or consequences of an activity in the domain of public programs. This definition is relatively open, and includes, for example, the prospect of both negative and positive consequences (Rieckin, 1952, p. 4).

The principal refinement of this rather broad definition is needed with respect to purpose. Independent of a consideration of actual methods or implications concerning rigor, the evaluation process can have either broad or narrow purposes. At the narrow end of the scale, the purpose may be thought of as specific to a very particular activity and involves supporting what amounts to a go/no-go decision on the part of a manager or administrator. At the broad end of the scale, the purpose is to develop generalizations about activities of a given type such that future policy level decisions can be enlightened. As a cue to this distinction, the process having a relatively narrow purpose could be labeled simply *evaluation* or *evaluative testing* while the process at the broad end of the scale could be labeled as *evaluative research*.

The process of evaluation can be viewed as differing from evaluative research in its focus on specific aspects of a particular program. Although both processes may share characteristics of common methodology (system design, data collection, and analytic procedures), the objective of research is to achieve knowledge that is highly generalizable, or to test hypotheses; the objective of evaluation is to provide a basis for selecting among alternatives. Research is primarily concerned with issues of relationships among variables, with assessing the effect of parameters common to many programs. Evaluation is primarily concerned with questions of practical utility that usually involve value judgments.

The present discussion is focused, therefore, upon the narrow zone of the purpose-of-evaluation dimension. It should be emphasized that this focus cannot permit any derogation of methodological rigor, for even narrow-purpose assessment is a complex process.

Critical Facets of Evaluation

Setting Objectives

A basic task in performing any evaluation is to establish measurable objectives. An objective is measurable to the extent that it becomes possible to state after evaluation the degree of success with which it was met. The primary tasks during the initial portion of developing a program evaluation plan are (1) to identify measurable objectives, (2) to identify associated elements of training programs that require assessment in terms of those objectives, and (3) to determine the methods and techniques of measurement which will best produce sound and usable evaluation data.

A first step is, therefore, to conduct a thorough review of the program itself. This activity often helps to achieve greater specificity of program objectives since it frequently reveals less than unanimous agreement among key personnel about program goals. Consequently, it is important to develop a written statement of program objectives—as agreed upon by participants at various operating levels. These objectives then serve as standards against which the evaluation is conducted. Where possible, it is desirable to formulate these objectives in measurable terms. For example, rather than indicating the desirability of involvement at the local level in retraining activities, it is preferable to set as an objective a specified level of involvement by a particular date. Thus, a standard is established against which program performance can be compared.

Outcome Analysis. Evaluation is an ends-oriented process. The critical question, i.e., the first question the evaluator must answer is: What is the end that the activity to be tested is set to accomplish? What, in a word, are the objectives? This point is so crucial that professional analysts are now declaiming that unless the administrative agency can specify precise performance objectives, there is no program. The proposition proceeds to the point of identifying "non-programs" versus "programs" on the basis of the specificity of performance objectives. (See, for example, Kidd et al., 1970, p. IV-6.)

In a sense, then, evaluation provides an instance of reverse thinking: the end is considered first, the means are considered second. The ideal specification of objectives is in behavioral terms (across the board, but particularly with respect to training programs). That is, the manifestation of the objective is in the behavior of a clientele or in the capability to behave in a way which is different as a result of the program or activity being evaluated.

Barring behavioral specification, the objective should at least be specifiable in terms of criterion dimensions. The implication is that a *change* in conditions is sought and that such change is discernible only if it occurs with respect to a criterion. We are clearly in the domain of relative or relational phenomena. Also, it is implicit that something which can be stated as a dimension is more susceptible to unambiguous measurement.

To summarize, the problem of the evaluator is really a series of problems. First, he must have (or formulate) the objectives in some version of an operational definition. Next, he must have criteria; that is, a specification of what constitutes the achievement of the objectives (e.g., change of such-and-such a magnitude in such-and-such direction) or what constitutes significant progress toward the objectives. Third, he must formulate or devise measurement procedures and the instruments that are needed to assess the procedures.

With all this, the evaluator may still have only the bare bones of an evaluative capability. Complexity is added by the prospect that the objectives may not be unitary. That is, it is more often the case than not that a given activity or program has multiple objectives. This condition opens up the additional prospect that the objectives sought are at least partially independent: achieving one does not ensure that the others are being achieved. Even worse, the set of objectives may include some that are contradictory: the heightened achievement of one objective can lead inevitably to the diminished achievement of another objective. The complete specification of the ends sought can require a complicated delineation of weights and priorities or trade-off conditions whereby the criteria are brought into some balance. In brief, the evaluator must often consider what is acceptable as well as what is desirable.

Side effects, whether anticipated or not, may be beneficial or detrimental. The point is that the analysis of objectives does not necessarily encompass all the effects or consequences of an activity which could be important. The evaluator may not be in the position to plan in advance with respect to criterial specification or measurement of effects outside those sought for as objectives but he should be in the position to observe and note such effects as they emerge because these "incidental" consequences can have value in both negative and positive ways.

Finally, the evaluator must be able to handle the cost factor. It is a bit simplistic, perhaps, to assert at this point that no activity is cost free. It is more of a contribution to recognize that cost can be considered as an outcome factor. It is even more to the point to consider cost as a complex (i.e., composite) variable and to differentiate the more gross subfactors such as startup costs, operating costs, and incidental costs. A complete outcome analysis would also include a consideration of intangible costs (e.g., stresses faced by program participants).

We can return now to our initial definition which included the term "results or consequences" and terminate our discussion on that phrase. Results, in the framework of operational evaluation, can be completely and adequately assessed only if the following conditions are met:

- Objectives are specified in behavioral performance or operational terms.
- Criteria are defined.
- Priorities are established.
- Cost factors are included.
- Measurement procedures and tools are available.
- Provision is made for detecting side effects.

Means Analysis. At the broad end of the purpose dimension of evaluation, the decision maker is often faced with a myriad of alternatives. His job could be characterized as the selection of the best means to get a job done. The prospect in such instances is a complete *comparative* evaluation of all candidate means as complete configurations.

In the middle range of the broad-to-narrow purpose dimension, the evaluative task is analogous to diagnosis and repair. The evaluation process in the middle range is likely to be focused on the comparative evaluation of components rather than total configurations. One of the implicit assumptions of the evaluative enterprise is that if a component is "good" (i.e., if it functions properly), it will make a positive contribution to the whole program.

At the narrow end of the purpose dimension, the process becomes more absolutistic: either the activity is doing its job or it is not. However, if it is determined that it is not, there usually must be some alternative to abandoning the objectives by simply terminating the activity under evaluation. If there is no other "fall-back" position, the program director must be prepared to use the evaluation findings as an aid to his only course of action which is marginal remediation.

This means that even in the go/no-go test situation, the evaluation should be set up so that any differential effectiveness of the component parts can be detected. The obvious further implication is that a scheme of component part identification be created and that, prior to the empirical phase of the evaluation, the interrelations or interdependencies of the component parts are determined. Ideally, then, the evaluation findings will automatically lead to identification of the weak links in the activity and bring the program manager to the threshold of corrective action. In a formal sense, all that can be expected of evaluative testing is that the principal contributors to any deficiency can be located. Evaluation findings cannot be expected to provide the precise prescription for remediation.

Constraints Analysis. No program or activity can be undertaken in a context free of constraints. The most prevalent constraint (and probably a universal one) is fiscal: there simply may not be enough money to do the job by means that are known to be most effective but which are marginally more costly. The next most prevalent restraint is time: the world simply will not wait for an extended planning and preparation phase.

There are more subtle constraints, however. For example, a program may require an assembly of personnel having a particular distribution of talent and capabilities and the right people can simply not be available. Essentially the same condition can exist with respect to equipment and facilities. In such cases, no conceivable amount of time and money could overcome such limitations.

Even more subtle are constraints imposed by the organizational or institutional setting within which an activity is launched. Organizations are not only rigid in the bureaucratic sense but limit program options because of a prevailing folklore or traditional belief about what is acceptable, feasible, or even "proper."

It is widely recognized that no amount of objectivity or scientific rigor in the evaluation process can exempt the enterprise from political considerations. Adverse findings are not always welcomed by program participants or even by the administrative instigator of the evaluation. The evaluator must face the prospect that the direct implications of his findings will be obscured by reationalizations in which the constraints on program operation are given great significance. A useful preventive to subversion of an evaluation effort entered into in good faith is for the evaluator to bring the constraints into the open at the outset and to realistically incorporate the consideration of the influence of these constraints into the report of the observations and measurements.

Focus of Evaluative Intervention

Historically, one of the practical problems in the conduct of public program evaluation was that evaluation was an afterthought and conceived as an adjunct to the basic job of program management. In some cases, in effect, the diagnostician was called in after the patient had died.

Modern practice is increasingly directed toward spending some effort in planning before an activity is launched and toward including evaluation in the planning agenda. There is still a choice, however, about the time phasing of evaluative intervention. No pat formula is available for making this choice but the alternatives can be examined with profit.

Basically, there are two options: during or after. In other words, evaluation can be an integral part of the ongoing activity or it can be intermittent such that it takes place at the natural termination of the activity or at the completion of well defined phases or stages. The first choice is generally labeled as *formative* evaluation while the option is called *summative* evaluation.

Formative evaluation refers to the process of obtaining and utilizing information that can be used to modify and improve a program. It is evaluation as a procedure for gathering and analyzing data on a program in progress such that results leads to improvements in the program which is then reevaluated. Formative evaluation is continuous. It functions to optimize program design through iterative feedback.

Summative evaluation refers to the terminal assessment of a finished product. It is aimed at providing information useful for making a *general* administrative decision about the program. In terms of a school system, summative evaluation might provide a basis for decisions about adopting a particular curriculum, or determining its effective use. In terms of the Belmont training seminars, summative evaluation would provide information on who participated, what and how much they learned, the effect that what they learned had within their agencies, and the cost effectiveness of the training procedure–in short, how effective was the program.

There are some good and bad features associated with both options. For example, formative evaluation has the advantage of early diagnosis. Investment costs can be minimized by terminating inappropriate activities ahead of schedule: failures can be aborted before they become monsters. Alternatively, corrective action can be taken before a failure reaches catastrophic proportions.

On the negative side, formative evaluation incurs the risk of premature judgment and/or obtrusive interference in program functioning. At best, formative evaluation is the source of some minimal level of distraction for program participants. In some cases, the provision of evaluative data can become a significant burden and then the evaluation process itself constitutes a program restraint.

One of the major lines of methodological development in the continuing assessment of evaluation procedures is the invention of observational and data collection techniques which are as unobtrusive as possible.

Summative evaluation has advantages and disadvantages which are essentially the converse of those for formative evaluation. That is, summative evaluation can occur too late to provide anything more than academic significance. Unless the activity in question is to be repeated, any corrective adjustments which might come from the evaluation are more or less empty gestures. On the other hand, summative evaluation is not obtrusive and is not a distraction for participants. Moreover, summative evaluation has the signal advantage of being compatible with a holistic appraisal; the activity can be looked upon as a unitary, coherent event. In this same regard, the effects of the activity that are naturally delayed can be more easily considered if summative evaluation is the chosen mode. The reverberations of an activity—some of its longer term effects can be assessed.

Other Aspects of the Evaluation Process

Effort Versus Efficiency Versus Effectiveness. There are many ways in which an activity to be evaluated can fail to function. The evaluation process can be differentially tuned to detect one sort of failure more readily than others. A common pitfall in evaluation, however, is limiting the focus of evaluation to one area of potential failure in the belief that a single area is indicative of the whole. This error is often impelled by lack of resources or methodological tools so that it is expedient to accept a narrow focus and promote the observations which result as a completely adequate picture of a total activity.

For the sake of discussion, three major areas of activity failure can be delineated: effort, efficiency, and impact. With respect to effort, one can assert simply enough that unless some measurable amount of energy is being expended in an activity, the activity is not functioning. We have here an instance of a "necessary but not sufficient" condition with respect to outcomes. That is, if no effort is going into an activity, we can be sure that there will be no outcome; but the fact of energy input does not assure outcome. Other breakdowns can intervene.

The simplicity of the proposition regarding effort can obscure two important implications which are less simple. Observation of effort can be a very important aspect of the evaluation process. Program failure resulting from lack of effort is not an empty category. Most public agencies have at least one program which is "on the books" but which is actually being neither supported nor manned. Given that the prospect of failure through lack of effort is real, one of the important implications is that evaluation of this kind of failure is relatively cheap and simple. Secondly, if a failure at this level is verified, no further evaluation need be done.

Efficiency is the ratio of energy expended to work accomplished. For present purposes, we will restrict the use of the term to the assessment of the contribution of the working parts or components of an activity to the outcome of an activity as a whole. In other words, we are talking strictly about internal efficiency. (The expansion of the efficiency concept to include the whole activity and its ultimate objectives will be considered later under the rubric of benefit-cost analysis.) The most critical indices of efficiency or the lack thereof, for present purposes, are the detection of wasted effort, less than capacity utilization of costly components, and duplication of effort. Another critical aspect would be the detection of cross-purpose operations within the activity under evaluation.

Consideration of failure in the matter of efficiency is somewhat analogous to failure of effort. Efficiency is a necessary but not sufficient condition to guarantee the overall success of the activity in operation. That is, it is possible for an operation to be efficient and still fail in a functional sense. One can be very efficient and economical about doing the wrong thing.

Similarly, the observation and measurement of efficiency tends to be easier and simpler than measuring the ultimate impact of an activity upon its environment. Efficiency is an "internal" matter and the internal components of an activity are under more control than are externalities. Such control facilitates observation and measurement.

Measuring effectiveness is always problematic; and always crucial. If an activity is given some effort and is working efficiently, it still may not be effective. However, finding out whether it is or not may be nearly impossible. For example, an educational activity might have as its goal the

acquisition of knowledge by a student but the student's utilization of that knowledge might not take place for many years. The ultimate effectiveness of many activities thus cannot be determined until well after it is too late to do anything about the activity in question.

The recourse of the evaluator is to focus his primary attention on interim outcomes: effects which are external to the activity but only moderately distant in time and range. Those events which take place at the boundary of an activity with its external environment are significant in that they tend to be observable and are often measurable *indicators* of effectiveness. For example, probably the most widely used indicator of effectiveness is some form of "customer" acceptance measure. If an activity has a discernible target audience such as the participants in a training program, the attitudes of the members of that audience are crucial to ultimate success or failure of the program.

We go a step in the direction of ultimate effectiveness measurement if we can add behavioral criteria to attitudinal. That is, has the behavior of the target audience member *changed* as a consequence of the activity? Have new, intended behaviors appeared? Again, however, there may be no way of eliciting the behaviors in question in proximity to the activity being evaluated. It is in these matters that most of the unsolved problems of evaluation methodology exist.

Spurious Rigor. Evaluation has its roots in the so-called behavioral sciences. For many years, the emphasis in the methodological development of these disciplines was toward increasing rigor both in the matter of the logic of experimental design and in the matter of the precision of experimental control and measurement. The model for these developments was the research paradigm of the physical sciences.

More recently, the validity of the physical science model has been called into question and emphasis has been shifting more toward a concern for relevance. Obviously, the ideal state of affairs would be one in which the methods of behavioral science were both rigorous and productive of relevant (as opposed to trivial) findings. That ideal state may come into being some time in the future but for the moment workers in the both theory-oriented and applications-oriented research must contend with a form of trade-off between rigor and relevance. Such compromises make many people uncomfortable and there is a tendency among scientifically trained people (such as those

likely to be responsible for operational evaluations) to continue to overemphasize rigor at the expense of getting meaningful results in the sense of answering operational questions. This is not intended to be an apologia for "sloppy" research but a suggestion about priorities and the avoidance of the form of "tunnel vision" on the part of evaluators. The search is for a coming-to-terms with reality. These matters are cogently discussed in a classic paper by Sinaiko and Belden (1965) and more recently by Finn (1969).

The Context for Inquiry

As suggested in the preceding section, evaluation is an administrative or managerially initiated function which, ideally, is planned as an integral part of the planning of the activity to be evaluated. Ad hoc, tacked-on forms of evaluation are almost universally worthless. The implications of this proposition are strong: the activity to be evaluated becomes precisely analogous to the independent variable in an experiment in the behavioral sciences. The overall planning of the activity becomes analogous to the design of the experiment. However, the experiment is one that must take place in the real world and not in a laboratory. Consequently, its complexity is vastly increased, the capacity for rigorous control is diminished (see above), and the range of feasible options for the logical structure of the experiment is greatly restricted. By and large, the preeminent option is an experimental design of the before-and-after type in which some target conditions are measured before an activity is launched and measured again after the activity is underway or after it is completed. The measure of effectiveness is the change in conditions between the two measurements. For example, with particular reference to Belmont training programs, participant attitudes toward the Belmont System before and after the training seminar, and the extent to which the seminar influenced those attitudes in positive or negative directions should be assessed. Assessment in this area would involve having participants characterize their attitudes toward the Belmont System (1) just prior to, (2) immediately after, and (3) some time after the seminar. Attitudes could be characterized in terms of rating scales, and the results quantified. The extent of positive or negative change could then be used to draw conclusions about the success of the program.

A second criterion of the success of the seminar might include Belmont-related activities with which participants were involved prior to, and subsequent to the seminar. Activities might include (1) completing, reviewing, consulting on, or supervising the completion of forms; (2) implementing, coordinating, or monitoring data collection procedures; (3) conducting workshops; or (4) serving as a resource person for a particular instrument.

The kind and number of personnel involved in these activities on a preseminar/postseminar basis should also be assessed. In making inferences about the operational level within State or local education agencies, that is absorbing the thrust of Belmont, it would be important to determine how many accounting personnel as opposed to instructional or administrative personnel participated in training activities.

The weakness in pure before-and-after experimental designs is in the lack of control over incidental interventions and processes that are going on simultaneously with the initiation of the activity in question. One cannot, on the basis of the results of a pure before-and-after experiment, unequivocally assert that the activity in question was the unique cause of the changes (if any) which were detected.

Many evaluations are, as a result of circumstances, carried out in the context of even weaker design situations. Most particularly, the measurement of the conditions prior to the initiation of the activity under evaluation cannot be made or are not made because of time pressures, political, or economic restraints. The consequence is a measurement situation which is labeled as an after-only design. The inferential logic for drawing evaluative conclusions in such a context must be based on presumptive scale-values assigned to conditions prior to—or in the absence of—the activity under evaluation. Fortunately, such presumptive assignments can be quite valid in many situations. For example, for a training activity in which trainees are to be taught a novel technique, it is reasonable to assume that their prior performance in the execution of the technique would be nil. There are also many circumstances where the outcome can be assessed against so-called normative levels of performance or against circumstantially derived but absolute standards. For example, the wider context of a training activity might dictate the requirement that subsequent to training, all trainees should be capable of performing a given task in a particular time with a predictable error rate. Effectiveness of the activity would then be judged against a criterion of the extent to which these standards were actually achieved. In such circumstances, an after-only form is both logical and economical.

—In some rare instances, it is possible to arrange the conduct of training activity such that the format of one of the more powerful experimental designs can be followed. A great increment in

inferential power can be achieved, for example, if provision can be made for control group comparisons. In designs that incorporate control groups, the influence of incidental and concomitant events can be detected and, in effect, deleted from the observations. A complete discussion of the logic of the arrangement of conditions for the conduct of evaluational inquiries in experimental design terms can be found in Campbell (1963) and most cogently in an article by Scriven (1967).

In summary, evaluation takes place in the context of a sort of quasi-experimental design. The format of the design and the consequent inherent power of the evaluation are largely determined by administrative (rather than scientific) considerations.

If the issues of concern in the evaluation process are those of administrative or managerial effectiveness, the contextual question is opened up even further. While the concept of constraint is recognized, the broader environmental context contains financial and political elements which quite often appear insignificant at the operational level but which nonetheless exert a potent influence on the character of the program. An awareness of and willingness to include such elements in an overall assessment of program effectiveness can often help resolve issues imposed by differing outcomes from apparently similar program efforts. Data collection efforts should, therefore, include provisions for obtaining such environmental information.

Specifying System Goals and Functions

In order to evaluate the effectiveness of, or to develop methods for evaluating any program, it is first necessary to have a clear understanding of what the program must *do*. It is always in terms of what a project does that ultimate evaluation or acceptance takes place. Thus, in the case of Belmont training, the training programs have been designed to accomplish several things and objectives have generally been focused on providing participants with: (1) a working knowledge of the Belmont System and selected portions of its developed instrumentation; (2) skills and techniques required to install the Belmont System within their own States; (3) materials required to mount training efforts within their States; and (4) an opportunity to gain insight into problems experienced with installation and operation of the Belmont System. Because the Belmont System is a continually changing, developmental system, it can be anticipated that future training will be necessary. It can

also be anticipated that future training will be conducted along the lines developed earlier. In such instances, one of the major outcomes of evaluation can be the retention of those features of the training activity that work well and the deletion of features which are shown to be inefficient or ineffective.

On a more general level, we can set the problems of evaluating Belmont training and similar training activities in an orderly framework. As in most such projects, there are layers within layers. Belmont training is part of a larger system of Belmont operations which is part of an even larger system of Federal support of local education. These layers are related to one another in much the same way that a larger, more encompassing system generates performance requirements for the smaller, subsumed systems. This situation allows for the establishment of a coherent sequence of dependencies as a framework for planning evaluation studies. The framework begins at the establishment of the demands or requirements that the larger system imposes on the smaller. Meeting those demands becomes the goal of the included system. Its objectives can then be stated in terms of the conditions of goal achievement with respect to extent of achievement, time, cost, and side-effect occurrence. The functions of the included system are its actions which are considered essential or conducive to the fulfillment of its objectives.

For example, the operations of an inclusive system depend upon the performance, by a designated group of people, of a particular job that the people in the group have not done before. The demand or requirement from the larger system is that the group of people be made capable of doing the job and that they be motivated to do the job in a conscientious and reliable manner without immediate coercions. The subsystem is then assigned the goal of acting upon these people in such a way that the requirement is met.

Objectives are set, in this instance, in terms of the numbers of people to be processed per unit time (i.e., rate of production) and quality standards (i.e., skill level at the completion of training, acceptable drop-out rates, etc.). The functions of the subsystem are the susceptible to delineation in the form of the content, mode, setting, and style of presentation of instructional messages or training materials (i.e., preparation of topical outlines, scheduling instructor-trainee contacts, specification of lecture versus discussion versus laboratory presentations and exercises, etc.).

Designing Evaluative Studies

The distinction between the contextual planning (overall quasi-experimental design) and planning the evaluation study within that context is important and one which escapes many persons who are in the evaluation business. The evaluation study proper involves only that measurement of outcomes that is feasible and relevant within the larger setting of the conduct of the activity to be evaluated. At this interior level, the same pattern of research logic obtains, however, and this is probably the source of some of the confusion. Thus, the program or activity to be evaluated is said to have objectives and the specification of these objectives is a precondition to evaluation. Then the evaluation study, itself, has its *own* objectives and these are the first priority consideration in the design of the evaluation study. The problem can be seen as a confusion of levels where the same terms are used across levels.

In general, the purpose or objective of an evaluation study is to provide information to the manager or administrator of the activity under evaluation. The design of the evaluation study is, then, partially determined by the uncertainty of the decision maker, the unresolved questions in the decision maker's mind, and the anticipated consequences of the decision.

The evaluator must take into consideration the needs of *his client* in his design of the evaluation study. It is up to the client to make the activity to be evaluated accessable to the evaluator. Evaluation that is initiated by sources that do not have both authority over and chartered responsibility for the activity to be evaluated is not likely to be successful.

The steps in designing an evaluative study are as follows:

1. Determining what to measure
2. Determining and scheduling data collection operations
3. Analyzing and interpreting the data
4. Reporting outcomes.

These steps are discussed in detail below.

Determining What to Measure. Although it is often difficult to specify in advance the precise nature of instruments to be used for measuring the degree of a program's success or failure, careful comparisons among stated program goals at various levels (OE/SEA/LEA-training staff-technical monitor-training committee members) will often suggest the kinds of techniques which can be developed or adapted to determine how well a program achieves its objectives. In many cases, an eclectic approach proves quite efficient and a variety of techniques may be employed. Belmont training provides a cogent example. Objective measures of participant knowledge about the Belmont System and its developed instrumentation might be taken on a pre- and postprogram basis. Items to be included in such a test could be as basic as identifying what the initials CPIR*, or ESS** stand for-or indicating the kind of data to be provided on various Belmont forms. Although in most cases, there will be a massive positive increment in the amount of information absorbed by the participants, a predetermined level of test performance could be set as indicating a successful training effort.

Particular care must be exercised in the area of content mastery, however, for seldom is there a clear trade-off between having available direct quantitative information on the degree of program success in the knowledge area on one hand, and, on the other, the probability of negative public relations caused by the possibility of embarrassing ostensibly knowledgeable individuals.

Another factor that bears on documenting the amount of learning that occurs by administering objective, content-oriented, examinations is the manner in which substantive material on the Belmont instruments is presented. Material presented via lectures, for example, generally can be expected to be short-lived in the memory of the participants.

In more general terms, training activity criteria can be placed in about four basic categories:

- knowledge acquired; factual or conceptual

*Consolidated Program Information Report.
**Elementary School Survey.

- skills acquired
- attitudes acquired–toward the job for which training was initiated
- attitudes acquired–toward the training process, as such.

All but the second of these criteria lend themselves to measurement through the use of paper-and-pencil instruments of the sort that are well established in education and psychology. The measurement of skill acquisition requires an actual performance of the skill in question; not a verbal description thereof. Consequently, the evaluation of skill usually involves the conduct of an exercise under controlled conditions wherein the individual trainee can demonstrate the skills in question to objective observers. Such exercises are inherently more expensive and time consuming than conventional measures of knowledge or attitudes and therefore may be considered of marginal feasibility even in situations where the overall activity objectives are primarily in the skill acquisition area. Under any circumstances, skill measurement operations must usually be tailored to the specific activity and the operations must be pretested for validity and reliability (see following section on Methodology for details).

Determining and Scheduling Data Collection Operations. Given a set of instruments and/or other observational techniques, the evaluator is faced with what amounts to a logistical problem. He must take considerable pains to arrange that the instruments, observers, other apparatus (if any), and the persons or events to be observed come together at a time and place scheduled in advance to coincide with particular phases of the activity under evaluation. In so-called real-world studies, which include evaluations, the timing is critical for two reasons: first, mis-timed evaluation can be a distraction to the main operations of the activity under evaluation; second, critical events tend to be unique–once they have occurred they do not occur again. In this same vein, premature intervention can lead to invalid or inconsequential findings which are at worst misleading to the decision maker and at best a waste of time and effort.

Analyzing and Interpreting the Data. The most common pitfall in data analysis for evaluation studies is the mis-application of statistical tests. The error is usually in the form of using tests which are based on assumptions concerning the scalar qualities of the measurement dimensions (i.e., ordinal

versus interval scaling) and the characteristic distribution function of the data (i.e., normal curve versus all other distributional forms). The motivation behind the disposition to make these kinds of errors is implicit in the proposition that if the assumptions are met, more precise and powerful discriminations can be made. The whole tendency is mistaken, however, not just in the sense that tests based on faulty assumptions can lead to erroneous conclusions but also because the nature of the problem usually does not require that precision and power inherent in the so-called parametric tests.

The issue is made clear by a consideration of the distinction between statistical significance and practical significance. It does not pay to make extremely fine discriminations which are significant in the statistical sense if the differences so established are too small to be of practical significance. Managers of public program activities are rarely concerned with the minutea that a six-factor analysis of variance test can turn up.

A second common pitfall occurs in the interpretation of data. The error here is in the attribution of causal influence. The rather sophomoric mistake of attributing cause to one factor in a correlation is still more common than most logicians would care to admit, but the problem is broader.

Part of the problem has its roots in the quasi-experimental structure of the larger context of activity initiation. This structure can be, at a superficial level similar enough to a true experiment that those directly involved are inclined to permit themselves the license of assuming that the overall process *is* an experiment.

The most subtle aspect, however, relates to the area of multiple and partial causation. The positive idea that evaluators should adhere to is carried in the term, influence. Most of the interventions, activities, projects, etc. which they are called upon to evaluate only *influence* the outcomes of concern to a more or less marginal degree. In the complexity of public programs, the delineation of a unitary cause in any process would be a historic event.

In summary, it can be said that in general the techniques for analyzing and interpreting data which were developed or refined and adapted in the behavioral sciences provide a rich resource for

application in evaluation studies but that these tools and the logic that accompanies them should be used with caution and discretion by the evaluation specialist. He should not be blinded by the apparent facility with which these tools can be used in laboratory settings. He should see both the residual limitations of these tools and, more important, the limitations inherent in the kind of data he is able to collect and the uses to which his findings will be put.

Reporting Outcomes. Every communication has a persuasive intent. The problem is who is to be persuaded about what. It is not the purpose (or should not be) of the evaluation study to persuade a manager that his activity is good, bad, or indifferent. The job of the evaluator is to persuade his client that the results of the evaluation are fair and accurate.

Problems of format and writing style are dealt with in detail in many handbooks and report-writing guides, several of which are precisely targeted for the matters at hand.* The only point which deserves to be reemphasized is that related to the presentation of the methodology and logic of interpretation that will fulfill the need to persuade the client of the fairness and accuracy of the evaluation.

Because the methods and logic are largely borrowed from the social and behavioral sciences (and given the likelihood that the evaluator has received some fundamental training in one of these disciplines), it is highly probable that the rationale and procedures used in the evaluation study will be couched in the jargon of the derivative fields. While such jargon is functional within the fraternity of behavioral sciences, it can be highly disfunctional in boundary-crossing messages. It turns out that the basic logic of test and evaluation is independent of terminology and that the description of procedures can be cast in ordinary language without adding unduly to the length of the discourse. Evaluators should present their findings in the language of the primary audience for the evaluation study; not in the language of their research colleagues.

*See for example U. S. Department of Health, Education, and Welfare; Office of Education. *Preparing evaluation reports*: A guide to authors. U.S.G.P.O., Washington, D.C., 1970. (Cat. No. HE 5.210:10065).

Methodology

General

In a sense, all the prior discussion has been concerned with methodology; but it has been aimed at what might be called the strategic-level problems. In this section, an attempt is made to get at more detailed problems at what might be called the tactical level in the conduct of evaluation studies.

Developing Instruments

Scriven (1967) has presented an expanded and slightly variant version of the four critical categories that were introduced in the preceding section. These are matched with a set of indicative behaviors which are reproduced in Table 1 in somewhat edited form.

Within each behavioral category, use of a particular type of instrument is implied. For example, discrimination (I.B.) is most economically and readily tested by means of some form of multiple-choice items. In contrast, analyzing and synthesizing behavior (II.A. & B.) is most conveniently assessed by some form of essay-type item.

For the evaluation of educational activities, generally, there exists a vast array of standard instruments for which score norms, administration procedures, reliability, etc. are established (see, for example, *The Buros Mental Measurement Yearbooks*, from 1949 forward). However, special programs usually require special instruments. This often means that the format and structure of existing instruments can be used but that the specific content of individual items must be revised to fit a particular situation.

When such instruments are tailor-made, they should be pretested for comprehensibility, administratability (ease of administration and scoring), reliability (consistency of score) and, hopefully, validity (although in many instances validity may not be testable because of the absence of ultimate criteria).

Table 1

Instrumentation Guidelines

Educational Objectives	Behavioral Manifestations
I. Knowledge	I. Knowledge
A. Items of specific information	A. Recital
B. Patterns of relationships, categorical knowledge	B. Discrimination
	C. Completion
	D. Labeling
II. Comprehension	II. Comprehension
A. Internal relationships, patterns of influence and interaction	A. Analyzing
	B. Synthesizing
B. Application and applicability of concepts	C. Appraisal
	D. Problem-solving
III. Motivation	III. Motivation
A. Broad, with respect to the area	A. Rating
B. Narrow, with respect to course content	B. Projection
C. Deep, with respect to learning	
D. Shallow, with respect to course	
IV. Nonmental Abilities	IV. Nonmental Abilities
A. Perceptual	A. Detection
B. Motor	B. Manipulation
C. Social	C. Demonstration

In any case, many problems can be avoided by the use of a gains-score procedure. This means that whenever possible, trainees should be tested prior to training on alternate forms of the instruments that are to be used for outcome testing. The gain between before and after scores is a measure which is minimally contaminated by biasing effects of individual differences and other incidental sources of score variation.

Design Tactics

A virtual checklist for the detailed design of evaluation studies is available in a recent series of articles edited by Stufflebeam (1968). Some of that material is excerpted with minor editorial revision below. The main thread is presented in outline form with a short preamble as follows:

> The logical structure of evaluation design is the same for all types of evaluation, whether context, input, process or product evaluation. The parts, briefly, are as follows:

A. Focusing the Evaluation

1. Identify the major level(s) of decision-making to be served, e.g., local, state, or national.
2. For each level of decision-making, project the decision situations to be served and describe each one in terms of its locus, focus, timing, and composition of alternatives.
3. Define criteria for each decision situation by specifying variables for measurement and standards for use in the judgment of alternatives.
4. Define policies within which the evaluation must operate.

B. Collection of Information

1. Specify the source of the information to be collected.
2. Specify the instruments and methods for collecting the needed information.
3. Specify the sampling procedure to be employed.

4. Specify the conditions and schedule for information collection.

C. Organization of Information

1. Specify a format for the information which is to be collected.
2. Specify a means for coding, organizing, storing, and retrieving information.

D. Analysis of Information

1. Specify the analytical procedures to be employed.
2. Specify a means for performing the analysis.

E. Reporting of Information

1. Define the audiences for the evaluation reports.
2. Specify means for providing information to the audiences
3. Specify the format for the evaluation reports and/or reporting sessions.
4. Schedule the reporting of information.

F. Administation of the Evaluation

1. Summarize the evaluation schedule.
2. Define staff and resource requirements and plans for meeting these requirements.
3. Specify means for meeting policy requirements for conduct of the evaluation.
4. Evaluate the potential of the evaluation design for providing information which is valid, reliable, credible, timely and pervasive.
5. Specify and schedule means for periodic updating of the evaluation design.
6. Provide a budget for the total evaluation program.

The above outline has been expanded by Worthen (1968) as indicated in Table 2.

Table 2

A Partial Classification Scheme of Evaluation Designs

Structure for Developing Evaluation Design	Type of Evaluation			
	Context Evaluation	Input Evaluation	Process Evaluation	Product Evaluation
A. Focusing the Evaluation				
1. Identify levels of decision-making				
2. Project and describe the decision situations				
3. Define criteria for each decision situation	w	x	y	z
4. Define policies				
B. Collection of Information				
1. Specify the source				
2. Specify the instruments and methods				
.				
.				
.				
etc.				

The format is then provided with commentary by Worthen (1968) as follows:

1. *Planning and focusing an evaluation*: Planning and focusing an evaluation is often performed by evaluation specialists in cooperation with administrators responsible for the planning and operation of the project to undergo evaluation. Tasks associated with this function are:

- Establishing premises which will guide the evaluation
- Determining what is to be evaluated and in what sequence
- Identifying the decision-making process as it operates in a given setting
- Identifying the decision-makers to be served
- Projecting the decision situations to be served
- Making explicit and clarifying project assumptions and criteria for each decision situation
- Learning the project objectives and operational procedures
- Restructuring, when necessary, the objectives into measurable behavior
- Determining the audiences and estimated deadlines for evaluation reports
- Reviewing the research literature concerning similar projects
- Using effectively subject area or technical specialists as consultants whenever necessary to review the evaluation plans
- Defining the staff and resource requirements for the specific project to be evaluated
- Constructing an evaluation budget for the project and securing necessary resources for evaluating the project
- Scheduling evaluation activities.

2. *Selecting or constructing instruments*: The next step is to select appropriate evaluation instruments or construct new instruments when existing ones are unsuitable for a particular situation. This involves the following tasks:

- Starting the purposes for which the instruments are to be used
- Developing criteria for selecting available instruments and selecting the most suitable instruments
- Developing specifications for constructing an instrument if no existing instruments are appropriate
- Developing, pilot testing, and revising new instruments.

3. *Collecting data*: The evaluator is now ready to administer these instruments in order to collect data judged in the planning and focusing stage to be relevant to the decision-makers' needs. Tasks related to collecting data are:

- Specifying information needs clearly and concisely
- Identifying information sources for collecting the data
- Specifying methods to be used in collecting data
- Specifying sampling procedures
- Specifying the schedule for data collection
- Training personnel to collect data
- Administering evaluation instruments and recording the data.

4. *Processing data*: The planning aspects of data processing occur simultaneously with those of instrument selection or construction and data collection so that after the data have been collected they are in a convenient form for processing. The processing function consists chiefly of scoring tests and other instruments and providing for data storage and handling. Tasks associated with data processing are:

- Providing a format for coding data
- Scoring instruments

- Providing for data storage, management and retrieval
- Coordinating data processing activites with other units within and outside the agency
- Using existing computer programs
- Writing new computer programs when necessary.

5. *Analyzing and interpreting information*: The raw data, after being processed, is then ready for statistical analysis in line with the evaluation design being employed, the nature of the data collected, and the level of sophistication required by the decision-makers who are to receive the information. The value judgments made in interpreting the information may be made by the evaluator himself or by expert consultants sometimes used for this purpose. Tasks related to analyzing and interpreting information are:

- Selecting the analytical procedures
- Designating a means for performing the analysis
- Performing the statistical computations
- Producing computational documentation when appropriate
- Interpreting the results of the evaluation program in terms of given criteria.

6. *Reporting Information*: This function serves to provide decision-makers with timely information that is relevant to their needs. Successful reporting of information is closely associated with an understanding of the (1) decision-makers' information needs, (2 characteristics of the audience to receive the information, and (3) estimated report deadlines. Tasks considered important for reporting information are:

- Specifying means for providing information to the audiences
- Specifying the format for evaluation reports
- Scheduling the reporting of information

- Providing evaluation abstracts or summaries for presentation to specific groups
- Preparing findings and recommendations to the decision-makers in an understandable manner
- Obtaining the decision-makers' reactions to the report.

Role of the Evaluator

A final problem to be considered is the personnel question. The basic issue in this problem area is whether evaluation should be conducted by personnel whose primary activities are intrinsic or extrinsic to operational aspects of the program—in short, who should evaluate what. Although there is no clearcut answer to the question as to who should conduct an evaluation, generally, evaluations concerned with improving detailed procedures and processes of the operational training program should probably be conducted by or based on data collected from project staff, and participants. Because these personnel function on a daily basis within the program, they are in the best position to assess the feasibility of suggested changes and improvements. Final evaluation of the cost effectiveness or summative evaluation of the overall program, on the other hand, should probably be conducted (or at least coordinated) by someone not closely involved with day-to-day operations. Some advantages and disadvantages of the two approaches are indicated below.

External Evaluator

Advantages

1. Likely to be objective
2. Unlikely to be distracted by operational problems
3. Able to concentrate full effort on assessment.

Disadvantages

1. Incapable of intimately understanding program
2. Ties up time of operational staff learning about program
3. Likely to interfere with operations by imposing perturbing measurement activities

4. External value structure imposed on project purposes
5. Uses funds better spent on refining optional aspects of program
6. May cause threat and resentment among operational staff.

Internal Evaluator

Advantages

1. Fully cognizant of all aspects of program
2. Not a disruptive influence
3. Inexpensive

Disadvantages

1. Lacking in objectivity and perspective
2. Ego-involvement will produce biases
3. Operational involvement will be at expense of evaluation

In many cases, however, there is no firm requirement for an either-or decision. A comprehensive evaluative effort will involve data collected from various sources, and the major decision will exist only with respect to who will integrate and interpret the data. In this area, an attempt might be made to counterbalance the advantages and disadvantages of internal versus external evaluation by relying more heavily on data collected from internal sources for formulative evaluation and more heavily on external sources for summative evaluation. The former approach capitalizes on the intimate knowledge of the program which its participants have and seeks to use that knowledge to optimize the configuration of the program. The latter approach capitalizes on the greater objectivity of the external data source.

Regardless of the approach selected, the role of the evaluator will vary depending on the level of development of the program, and stage of the program at which his inputs are desirable. Some appropriate evaluator activities during various phases of the program are suggested below.

1) Activity During the Formulation Phase

During the program formulation period, the evaluator's major role should be concerned with formulating criteria for assessing agreed upon objectives. The requirement for evaluation should be embedded in the program design in a way that causes it to be viewed as an integral and accepted aspect of program operations rather than a process that was added as an afterthought. The notion of evaluation as an appendage practically guarantees potential conflict with program operations.

2) Activity During Development and Implementation Phases

In an ideal sense, the evaluator should be available to assist the program developer to establish clear and realistic distinctions between program goals and objectives during the development phase.

3) Activity During Operational Phase

During this phase, the evaluator should concentrate on providing inputs to the program managers, on the basis of which program modifications can be implemented. Modifications can occur with regard to any aspect of the program, including goals or objectives.

4) Activity During Postproject Phases

During this phase, the evaluator's primary function is to accurately describe the program, and to determine the relationship between program goals and outputs, that is, to assess how successful the program was in meeting its objectives. The evaluation product should thus describe program outcomes in quantitative terms as well as assess the value of those outcomes. (It should not only provide a detailed answer to the question, *What happened*? but also the question, *How good was it*?) Information should be provided to suggest methods and techniques for improving subsequent programs and the inferences about outcomes which may be drawn from continuing the program along its present lines. Areas of strength and deficiency should be clearly identified, and information provided for use in planning future efforts.

To be particularly avoided, however, is the approach which involves the development of a series of general impressions about program adequacy by expert participant-observers. Frequently, the suggestions which result are obtained on the basis of very limited exposure to the operational

context, and severely limited data gathering. In many instances, the nature of the recommendations are more reflective of observer biases than any quality inherent in the program. While it may seem advantageous to encourage this approach to assessment during the formative stages of program development in the interest of increasing program efficiency, the value of these efforts can be predicted to be quite limited. A nonsystematic approach to evaluation practically guarantees that "recommendations" will be contradictory, and that evaluation sessions will be characterized by disagreement and inconsistency. While the approach may have much to recommend it from a pragmatic standpoint, difficulties arise when these activities are viewed as adequate substitutes for systematic evaluation.

In contrast to the approach of using an array of participant "experts," the ideal evaluation team would consist of separate individuals fulfilling each of the following roles: (1) Evaluation Director, (2) Evaluation Coordinator, (3) Surveillance Specialsit; and the following support specialists: (4) Instrument Specialists, (5) Data Collection Specialist; (6) Data Processing Specialist, and (7) Reporting Specialist. The role requirements and task elements in each of these roles would be as follows:

Evaluation Director. The Director should have a background in research management and an understanding of both instructional operations and theoretical and practical aspects of evaluation. Specific responsibilities that could be listed in a job description for the Evaluation Director are:

Maintaining continuous contact with administrators, and project directors regarding evaluation needs

Coordinating evaluation activities

Identifying and reducing any inhibitions toward evaluation in administators and other instructional personnel

Disseminating the purposes and advantages of evaluation to instructional personnel

Directing the planning and focusing of the evaluation

Defining staff and resource requirements and planning for meeting these requirements

Constructing and managing a budget for the total evaluation program

Reviewing all evaluation designs, instruments, and reports before they are used or released for distribution.

Evaluation Coordinator. Such a person could be responsible for coordinating all evaluation activities related to the one or several projects to which he is assigned. He should become familiar with the substantive area of each assigned project as well as with the project directors and their staffs. When working with any project's support personnel, he should insure that they are familiar with any unique conditions of the project being evaluated that may influence their technical operations. When each support specialist has completed work on his project, the Evaluation Coordinator should be sure that he understands what has taken place so that he can adequately explain it to the decision-maker to whom he reports the evaluation. The specific tasks for this role are:

Maintaining frequent communication with and observation of the project director and participants

Interpreting the decision-making process as it operates in the project to be evaluated

Identifying the decision-makers and the decision situations to be served

Clarifying the project objectives, if necessary, and defining criteria and measurement techniques for each decision situation

Coordinating the scheduling and administering of data collection instruments to project participants

Defining the project systematically

Providing feedback of evaluation information to project participants as-well as information for interpreting and utilizing it

Obtaining the decision-makers' reactions to evaluation reports

Using the services of the support specialists in the evaluation unit whenever necessary.

NOTE: In most evaluation situations, it is recognized that this would be a low-priority role.

Surveillance Specialist. In education, the Surveillance Specialist could be a person who spends much time outside of his agency looking for new ideas. Job specifications for this role are:

Detecting promising innovations in other agencies

Scanning the research literature for ideas that might be tried

Presenting reports on innovative practices observed or discovered in the literature

Preparing abstracts describing innovative practices observed and disseminating these to instructional personnel.

Instrument Specialist. This specialist would have general responsibility for selecting and developing tests and other instruments to be used in evaluations. Specific tasks for this role are:

Selecting available instruments when appropriate to the needs of a given project

Analyzing the strengths and weakness of any instrument

Developing test plans

Writing, or supervising the writing of, acceptable items for instruments

Pilot testing new instruments

Making appropriate validity and reliability checks on instruments

Revising instruments and preparing directions for administering and scoring the instruments

Performing research related to new measurement techniques

Maintaining current information on newly developed instruments.

Data Collection Specialist. This role calls for technical knowledge regarding data collection techniques as well as supervisory skill in selecting, training and supervising para-professionals who actually gather the data. Tasks included are:

Interpreting information needs observed by the evaluation coordinator as they affect the data collection process

Determining from the evaluation coordinator existing and potential sources of information.

Specifying instruments and methods to be used in conjunction with the instrument specialist

Selecting sampling procedures

Specifying the conditions and the schedule for information collection

Determining the qualifications and training needs of personnel to collect the data

Employing and training personnel to collect the data or contracting out for this service

Coordinating with the data processing specialist in determinig a format for coding information

Performing research related to new sampling and data collection techniques.

Data Processing Specialist. A person performing this role needs competence in statistical analysis, computer operations and information management. He also needs skill in explaining these technical tasks to other evaluators and educators in terms they will understand. Like the data collection specialist, the Data Processing Specialist also needs supervisory skill in selecting, training and supervising para-professionals who are to key-punch, score tests and perform routine computations. A job description for this role includes:

Providing a format for coding information collected

Scoring and providing item analysis for instruments

Providing for data storage, management and retrieval

Coordinating data processing activities with other units within and outside the agency

Maintaining current information on new and existing computer programs and other data processing systems

Explaining, when necessary, computer operations and outputs to other evaluators and educators

Writing basic and intermediate level computer programs

Coordinating the writing of more complicated programs with an experienced programmer.

Selecting the analytical procedures and designating a means for performing the analysis

Interpreting the results in terms of given criteria.

Reporting Specialist. Personnel in business and industry have often been more aware than educators of a need to communicate ideas clearly and interestingly. The role of the Reporting

Specialist is primarily one of assisting evaluation coordinators in providing both oral and written feedback of evaluation information to persons both within and outside the school system. This person should also be familiar with audio-visual techniques for presenting information to individuals and to groups. Specific tasks for this role are.

Specifying the audiences to receive evaluation reports based upon information provided by the evaluation coordinator

Determining the commonalities and differences between the audiences in order to judge how many versions of the reports will be needed

Selecting a format for oral and written evaluation reports that provides relevant information concisely

Providing evaluative abstracts for presentation to specific groups

Preparing findings and recommendations to the decision-makers in a variety of interesting and understandable forms, such as: written reports, short films, filmstrips, video tapes, wall charts, overhead transparancies or other media

Obtaining the decision-makers' reactions to the reports in order to provide a basis for improving future reports.

Educational practicioners at the local, state and national levels are becoming increasingly aware of the need for personnel highly skilled in the theoretical and practical aspects of evaluation. Nevertheless, few educators recognize the heterogeneity of tasks required in an evaluation unit or agency. Such diversity of tasks has caused some writers to propose emerging roles for various specialists within an evaluation unit. While the variety of specialists within an evaluation is limited by the total number of personnel working in the unit, it is important that administrators explore the types of evaluation to be performed and select personnel with the various skills required to form a balanced and integrated team of specialists.*

*Much of this section was based on the article by T. R. Owens cited previously.

REFERENCES

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental design for research in teaching. In N.L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Finn, J. D. Institutionalization of evaluation. *Educational Technology*, 1969.

Kidd, J. S., Davis, R., and Gardner, R. Program analysis and urban priorities. In E. Winchester, and P. Finkel (Eds.), *The Role of Analysis in Establishing Program Priorities*. APPA Symposium Proceedings, Washington, D.C., May 1970.

Lumsdaine, A. A. (Ed.) *Evaluative research: Strategies and methods*. Pittsburgh: American Institutes for Research, 1970.

Owens, T. R. Suggested tasks and roles for evaluation specialists in education. *Educational Technology*, September 1968.

Reicken, H. W. The Volunteer work camp: A psychological evaluation. Cambridge, Mass. Addison-Wesley, 1952.

Scriven, M. The methodology of evaluation. In *Perspectives of curriculum evaluation*. AERA monograph series on curriculum evaluation. Chicago: Rand McNally, 1967.

Sinaiko, H. W. & Belden, T. G. The indelicate experiment. *Proceedings of the Second Congress on the Information System Sciences*. Hot Springs, Va., 1965.

Stufflebeam, D. L. Toward a science of educational evaluation. *Educational Technology*, July 1968.

Suchman, E. A. *Evaluative research*. New York: Russell Sage Foundation, 1967.

Worthen, B. R. Toward a taxonomy of evaluation designs. *Educational Technology*, August 1968.